Discussion

How did you preprocess this dataset ?

1.資料讀取進來後，首先應該看的應該是是否有NAN的欄位，若是有的話可以將該欄位剃除，不過在資料量不大的情況之下剃除欄位會造成資料更少，因此較佳的辦法是填補遺漏值。有幾種基本的填補方式一種是填平均值，但是這種方式會無法反映出資料的資訊(變異數降低)，另一種是用knn，不過我們的資料集並無缺失值因此可以安心使用。

2.outlier的剃除，照理來講這個剔除的應該依據盒鬚圖去做判斷的依據，不過python似乎沒向R那麼好使用，因此我這裡先用直方圖去判斷，發現大多數的Volume應該落在45\*10^8內，因此粗略將超出這個值得視為outlier。

3.將明日的畢盤價-今日畢盤價>0則為漲(1)<0則為跌(0)由此計算出之後所需要的y=next day raise。

4.將Date轉換成月份並且使用dummy variable轉乘啞變數這樣才不會造成12越是1月的這種無法解釋的數值關係。

5.資料分割:將train\_x, train\_y,test\_x,test\_y做分割x可以是自己選擇的參數，不過從heatmap去看相關係數可以發現各個參數對於next day raise的相關係數得非常接近0，感覺不是一個太好的現象

6.將train\_x與text\_x做正規化的動作，避免像是volume的直每個都超大因此影響了其他因素的重要性

Which classifier reaches the highest classification accuracy in this dataset ?

|  |  |
| --- | --- |
| Logistic Regressiom | 50.79% |
| SVM | 52.38% |
| NN | 47.62% |

SVM reaches the highest classification accuracy in this dataset

Why ?

這個問題十分難以回答，因為三者模型根據我預測出來的結果其實並沒有差很多，但我還是盡量去做解釋。

對於SVM與LR因為都是同為二分法，而且預測值其實跟0.5沒有太大的差距，因此兩者不太有”比較佳”的現象，但由於SVM對於outlier的影響較小，雖然說預處理時已經有粗略剃除離群值，不過並不是非常正規的做法，但這可能就是造成SVM正確率略高於LG的原因。再來，整體股票的趨勢其實是上漲的，next day raise的1的個數大於0的個數，因此用0.5當作一個threshold有點不公，因原先的資料就已經imbalance了，所以什麼情況都不知道的情形下去猜1的準確率還會比較高。

對於NN來說，資料及本身的影響較小，因為在backward-propagation的時候就會自行調整各項權重，資料探勘教到NN在解釋上不好解釋，因為就是參數丟進去然後結果出來，所以在解釋上比較無法向前兩者有合理的解釋。

Can this result remain if the dataset is different ?

我覺得這個答案應該是肯定的，不同的資料在套用模型上本來就會有所優劣，例如在LR比較適用於2分法的東西，多餘3種以上的分類SVM或許會有比較好的效果，至於NN比較大的影響要看輸入得資料與其參數是否配合得當。因此不同的資料集在不同的模型與參數調教之下會有不同的結果。

How did you improve your classifiers ?

要有所提升可以大致從兩方面去下手

A從資料方面去做選擇剃除

B從模型參數去做調整

1.除了上述的outlier剃除改用盒鬚圖可以比較準確之外，覺得有一個重要的因素是對於時間的描述並沒有考慮進去，股票會隨著時間軸的拉長而不斷上升，然而如果沒有將時間軸考慮進去test的price幾乎都比train的高，因此在分類上很容易直接將test的都歸類到會漲價。

2.使用feature selection做比較有效的參數選擇，我這次的feature幾乎都是全當參數丟進去，但從熱點圖可以看出有些參數的相關係數很高代表共線性高，因此將全部的feature都丟進去可能會將某項本身不是特別重要的feature重複計算好幾次，讓他的權重變重，因為共線性的原因。

3.參數的調較，這項對於NN尤其為重要，參數的選擇一直以來都是NN備受討論的話題，包刮hidden layer的數目......等，都是可以去做調教的部分，但這部分老師沒有深入討論我也只是利用助教給的參數去跑模型，事實上目前為止也沒有一個特定得方法去說參數一定要怎麼調整，大多數是依靠經驗以及trial and error去調整參數。